
Talking One Genetic Language:

The Need for a National Biotechnology Information Center

Biotechnology: New Genetic Miracles

Because of Biotechnology, a term coined less than five years ago:

- A man with diabetes is thriving on a new, life sustaining medicine. He had had a serious, potentially fatal, allergic reaction to various forms of insulin (traditionally extracted from a beef or pork pancreas).

Now he is being treated with pure human insulin, manufactured not in the body, but in the laboratory. To make it, scientists inserted "good" human insulin genes into fast growing bacteria, which then churned out great quantities of insulin hormone.

- New hope has come to a woman with hairy cell leukemia, a rare cancer. Her doctors are giving her infusions of alpha interferon, another new substance manufactured in the laboratory by splicing genes into bacteria. The results are amazing: the leukemia is simply disappearing.

And tumors in the bodies of over 50 patients—including lung and colon tumors notoriously resistant to conventional therapies—shrank by at least half when they were treated experimentally, with "interleukin 2." This genetically engineered substance turns the body's white blood cells—often in short supply—into specialized cancer killing cells.

- Thick mucous clogs the little boy's lungs. The doctors suspect cystic fibrosis but hesitate to start difficult therapies until they are sure of their diagnosis. Geneticists are called in to do a probe. By mixing a sample of the boy's blood with a piece of DNA (deoxyribonucleic acid) they determine that a crucial gene is missing. The test is positive; the diagnosis is made; antibiotic and other treatment can begin.

Biotechnology: Roots

New natural drugs and vaccines. Life saving therapies. Pain-saving interventions.

All these are the result of Biotechnology—research and development involving the all important molecules that control our life processes—how our bodies grow, how we age, whether we suffer a host of mental and physical diseases. Its central focus is DNA, the long, twisted threads in the nucleus of each of our ten trillion cells.

Scientists have known for some thirty years that genes are essentially pieces, or chemical subunits, of DNA carrying the messages of heredity in the strands of the famous double helix. A string of thousands of these units, which themselves are groups of atoms composed of four different nucleotide bases—A (adenine), T (thymine), C (cytosine), and G (guanine)—makes up a gene.

What matters is the arrangement of these chemicals, or their exact sequence in various combinations, along the backbone of the DNA molecule. For if the four different DNA bases are the letters of nature's alphabet, the chemicals' arrangement in groups of three is a sort of genetic code—dots and dashes that spell out all the instructions the body needs to manufacture the myriads of proteins which build our bones, our muscles, the enzymes which catalyze our metabolic processes, and all in all make us human beings marvelously different individuals.

If researchers can read and understand the language of heredity—if they can learn the sequence of bases in a gene—they can determine the makeup of the protein for which the gene is the blueprint. And they can clone that protein outside the body. Or find out why certain genes are switched on, or off during a lifetime. Or which ones are defective, or even missing.

When scientists first began translating genetic messages, they had a difficult time. But in the past few years, they have gained powerful new automated tools with which to decipher and analyze, or sequence the messages of heredity. And they have developed ways systematically to change DNA and other important molecules, that is to manufacture new genes and perhaps repair defective ones. As they have

done so the sciences of molecular biology and molecular genetics have become Biotechnology.

How Biotechnology Information is Unique

The complexity and size of Biotechnology information astound the imagination and make it different from other scientific information. And this information is growing today at an amazing rate.

Within every one of us, tens of thousands of individual human genes control special life processes. Three billion units of DNA make up the human genome (all human genes taken together), and only .01% of them have been sequenced. With current technology, sequencing these three billion units could consume 30,000 person-years and upward of \$2 billion. Fortunately, current technology does not stand still.

Technical advances—including ways to separate large molecules and individual chromosomes—have upped the rate of analysis about ten times in the last decade. A molecular biologist at a technically advanced laboratory today can sequence about 300 DNA units a day. And now scientists at CalTech have reported the successful development of an automatic DNA sequencing machine that again speeds sequencing tenfold (and cuts the cost for each base in half—from one dollar to fifty cents).

All this effort, all this information would be useless without computerization. There is no way even an Einstein could analyze, store or manage it with a pencil and a human brain. As information has poured out of the laboratories, factual research data bases have been developed to store it. There are now about a dozen such data bases, set up for the most part by researchers with an avocational interest in computers. The chart on the next page describes these data bases.

The Biotechnology Information Gap

The problem is simply stated but hard to solve. The data bases are swamped. Take the best known, GenBank, set up under contract by NIH and co-funded by a number of Federal agencies. By mid-1986, only 54% of the data published in 1985 had made it into the data base.

This is because the rate of publication has grown so rapidly—from one sequence composed of 76 bases in 1965 to a total of 11,552 sequences composed of 9,924,741 bases by 1986. The contents of GenBank by year of publication are in the table on page 4.

What is at stake here is not merely the convenience of researchers searching for scientific papers. What is at stake is the progress of Biotechnology around the world—our understanding of life at a molecular level and hence our knowledge of health and disease.

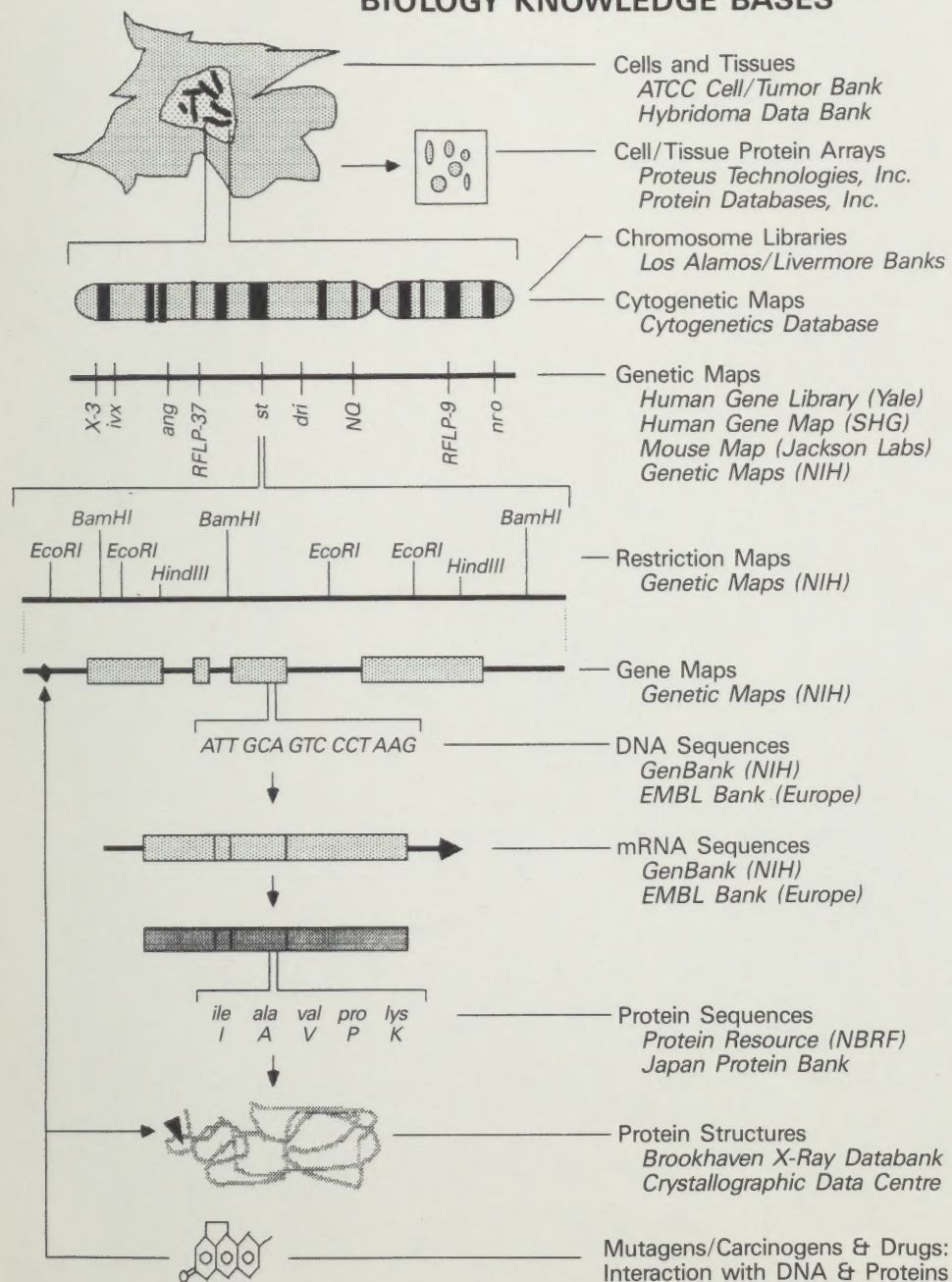
For scientists depend on the data bases actually to accomplish their research. They must tap into a base like GenBank to find matching sequences.

For example, they had known for a decade or so that oncogenes transform normal cells into cancer cells. Wondering why Nature would put the seeds of our own possible destruction within us they searched the data bases and discovered that the oncogene sequences matched those of a normal human growth gene. This suggests that cancer may be caused by a normal gene being switched on at the wrong time. And that they might some day find a way to switch it off.

But the different data bases use different information systems—different computer languages, and these have resulted in a veritable Tower of Babel. No scientist in the world can know enough about each of these computer systems to tap into all of them.

Even if such a genius appeared, the data bases lag so far behind that he or she might miss a discovery already made. And society might miss a new drug like Captoril, which controls high blood pressure—designed by knowing what a molecular target in the body looks like, and synthesizing a molecule that attaches to that target. Or a just-announced test that uses genetic material to detect the AIDS virus in blood samples.

BIOLOGY KNOWLEDGE BASES



Needed: a National Biotechnology Information Center

The whole biotechnological information system is so overloaded that there is a danger scientific progress may grind to a halt. To prevent that from happening, and to move the field along faster, we need a central repository for storing and sharing the information resulting from genetic research. With this problem in mind, Rep. Claude Pepper introduced a bill (H.R. 5271, 99th Congress; reintroduced in the 100th Congress as H.R. 393) to create a National Biotechnology Information Center, at the National Library of Medicine.

Working with the laboratories from which information comes, experts at such a center would seek to coordinate data as it is accumulated—to store, process and make it available to the research community, nationwide.

At such a center, too, computer scientists—with the help of some of the world's outstanding molecular researchers working next door at NIH—would create new computer information systems so that investigators throughout the country could ask questions and get answers quickly. They would encourage consistent terminology for researchers and data bases so that research results entered into computer systems could be shared and be made widely available.

In the new information systems, research would be stored in such a way that data retrieved from one source could be linked to other related findings, and to other research data bases. So investigators could ask one computer a question and that computer would automatically search for the answer not only in its own knowledge base but in other research bases as well.

Why an Information Center at the National Library of Medicine?

Developing such a national center at the Library would, of course, prevent duplication. It would enable scientists to work in a more collaborative style—conferring frequently to avoid reinventing the wheel.

Most important, it would give researchers the means they need to do their job. Now each of them is essentially alone, working on his or her own piece of the genome puzzle. What they need is something they cannot get in any regional or university center: a national “board” on which they can fit that incredibly complex puzzle together.

Starting a national center at the Library would be a cost effective, logical move. For it would build on the unquestioned leadership of the library in biomedical information. Nor would it be necessary to build a new facility, or to invest in extensive computer equipment to get fast results.

GenBank Contents by Selected Year of Publication
(as of August 1986)

Year	Bases	Sequences	Av Seq Len	% of Published
1965 *	76	1	76	
1970 *	249	3	83	
1975 *	6,160	54	114	
1980 *	439,717	852	516	90
1983	1,929,368	1,630	717	80
1984	2,569,480	2,475	1,038	75
1985	2,329,394	1,957	1,190	54
1986	290,708	161	1,806	4
Total (1965-1986)	9,924,741	11,552	859	82

This is because the Library is already deeply involved with the communication of biotechnology research findings, through its biomedical literature and computer data bases. Over 97% of all published research containing DNA sequences can already be found among the 6 million references in the MEDLINE (MEDical literature onLINE) system. The Library has twenty-two years of experience in building and maintaining large computer files for biomedical researchers and health care professionals.

Many of the tools and techniques developed by the Library to classify information and biomedical literature can be applied to the understanding of the language of molecules. For example, the

computer methods used to search for matching words and phrases in the medical literature can be applied directly to the finding of matching patterns of molecules in different DNA gene sequences.

The Payoffs: Toward Better Health and Less Disease

The functions proposed for the National Biotechnology Information Center seem to be absolutely necessary to do the genetic research job that has to be done in the years ahead.

We need such a Center to prevent duplication and to get more bang for our research bucks. We need it to fit

together the pieces of the genetic puzzle and acquire the knowledge that would benefit humankind in many ways.

With such knowledge, we could develop deeper understanding of the root causes of diseases that still plague us. This is true of diseases we know are inherited, of course, like Sickle Cell Anemia or Thalassemia.

But scientists feel that our likelihood of falling victim to the common and serious diseases, like heart disease, cancer, and arthritis, may well be influenced in complicated ways by the workings of our genes. Researchers aim now to learn more about these genes, and their relation to each other.

From this deeper understanding would come exciting new cures and interventions:

- New vaccines to prevent diseases and newer and more effective drugs to treat them;
- Gene therapy to cure even previously fatal disorders;

- Early warning for such diseases as Huntington's Chorea;
- Control of such dreaded conditions as Alzheimer's Disease;
- Better health generally through more plentiful food supplies—gained by inserting genes to make new plant species resistant to insects and herbicides.

Down the road a bit we can envision a scenario like this:

A single defective gene has left a little girl with a horrible immunological disorder called ADA (Adenosine Deaminase Deficiency). She is emaciated, and racked with pain. Because her body has absolutely no defenses against infection, she has had pneumonia almost all her life. Similar, luckier, patients have been helped by transfusions from a brother or sister's bone marrow that could be matched with their own. But only 3 out of 10 patients have such siblings, and this girl is not one of them.

Now this patient is about to receive an exciting form of gene therapy. Building on successful research results, the doctors plan to insert a normal gene into her bone marrow cells. They believe this treatment will produce an enzyme to correct her condition, and cure her.

Another scenario: A vigorous young man has developed a cancer of the lymph nodes. A genetic probe shows that the disease is due to the switching on of an oncogene in his body. Doctors inject a medicine containing a repressor molecule, and this switches the oncogene off. He recovers immediately.

Such miracles, such real cures, are expected from biotechnology research. This is the work the National Biotechnology Information Center is to assist.

